KENTUCKY DEPARTMENT OF EDUCATION

STAFF NOTE

Review Item:

Reliability and Validity Studies Update

Applicable Statute(s) or Regulation(s):

KRS 158.6453

History/Background:

Existing Policy. According to KRS 158.6453 (5), "…the Department of Education shall gather information to establish the validity of the assessment and accountability program. It shall develop a biennial plan for validation studies that shall include, but not be limited to, the consistency of student results across multiple measures, the congruence of school scores with documented improvements to instructional practice and the school learning environment, and the potential for all scores to yield fair, consistent, and accurate student performance level and school accountability decisions. Validation activities shall take place in a timely manner and shall include a review of the accuracy of scores assigned to students and schools, as well as of the testing materials. The plan shall be submitted to the Commission by July 1 of the first year of each biennium. A summary of the findings shall be submitted to the Legislative Research Commission by September 1 of the second year of the biennium."

The complete Scope of Work for the Office of Assessment and Accountability, Kentucky Department of Education (KDE), was reviewed by the National Technical Advisory Panel for Assessment and Accountability (NTAPAA) and by the Kentucky Board of Education, and was submitted to the Legislative Research Commission in August 2000.

Research activities described in the Scope of Work are undertaken to provide evidence as required by the statute quoted above. These research activities cover a broad range and, among others, include the following topics:

1. School Visit/Interview Study
2. Annual Multiple Assessments/Convergence Validity Evidence
3. Annual Third-Party Checking of KCCT Scaling and Equating
4. Annual Item Content, Item Difficulty, and Item-Type Mapping for the Multiple KCCT Forms
5. Biennial School Classification Accuracy
6. Annual Student Classification Accuracy Analysis
7. Technical Reports
8. Other Studies

A brief description of each study is given below. A more detailed description and update for each study is provided in Attachment A.

1. **School Visits Study.** Annual KDE validation-sponsored visits to Kentucky schools began in the 1996-97 school year for the purpose of linking the state's testing and accountability system to educational practices. Toward that end, the impact of the accountability system on teachers' instructional methodology, course content (or classroom curriculum) and teacher professional development have been examined, and successful programs have been documented.

2. **Annual Multiple Assessments/Convergence Validity Evidence.** It is helpful to explore the relationship of KCCT results to Kentucky student demographic data, CTBS/5 test scores at grades 3, 6, and 9, student questionnaire data, and ACT results for some high school students. These studies allow examination of convergent and discriminate relationships as well as differences in performance related to gender, ethnicity, and socio-economic status (SES). The student questionnaire data offer analyses of student characteristics (e.g. motivation, SES) in relation to test scores as well as students' perspectives on classroom instructional practices. The ACT data include students' self-reported transcripts, allowing analyses of course taking patterns and test performance. Any of these variables can be aggregated to the school level of analysis. Trends over time are also of interest. Knowledge of the relationship of KCCT scores to those of other assessments help support the valid interpretation of KCCT results.

3. **Annual Third-Party Checking of KCCT Scaling and Equating.** Psychometric analyses of KCCT data are conducted in parallel fashion by two independent professional organizations within the same operational time frame. This procedure ensures the accuracy of the computations involved in the main contractor's work. The two analytic groups operate in tandem, checking the agreement of intermediate results, noting and resolving discrepancies. HumRRO generates a yearly report detailing discrepancies and steps taken to address them.

4. **Annual Item Content, Item Difficulty, and Item Type Mapping for the Multiple KCCT Forms.** Item mapping is simply displaying two-way distributions of item content, difficulty, and type within and across the 6 or 12 test forms administered for each KCCT grade/subject combination. These maps summarize content validity and comparability of forms in terms of content and difficulty to address the following:

    1. Is the Kentucky Core Content for Assessment equitably represented across forms?
    2. Is the Kentucky Core Content for Assessment proportionally represented across Novice, Apprentice, Proficient, and Distinguished (NAPD) achievement levels?
    3. How is the Kentucky Core Content for Assessment represented by item format (multiple-choice versus open-response)?
    4. What is the distribution of item formats by Novice, Apprentice, Proficient, and Distinguished (NAPD)?

5. **Biennial School Classification Accuracy.** At the end of every CATS accountability cycle, Kentucky public schools are placed in one of three classifications (Meets Goal, Progressing or Assistance) defined by each School Growth Chart and based on end-of-cycle KCCT, NRT and non-academic indices. This collection of data should provide a very stable base for making classification decisions; however, because no measurement system is perfect, it is important to specifically document this accuracy.

6. **Annual Student Classification Analysis**. The Kentucky Core Content Test is administered annually in 18 different school levels by core content subject combinations (e.g., Grade 4 Reading, Grade 8 Mathematics). Based on their responses, students are classified into one of four basic categories (Novice, Apprentice, Proficient, and Distinguished, commonly referred to as NAPD) and the classification results are used to compute school accountability index scores. Given that no test is perfectly reliable, it is important to document the accuracy of these student classification decisions.

7. **Technical Reports (Reliability and Validity Evidence)**. At the end of each biennium, the Department, along with it assessment contractors, produce a Technical Report documenting many facets of the KCCT, including item and forms development, scaling, scoring and reporting, and evidence for reliability and validity. In odd years, a set of Technical Appendices is created instead of a full report. The tables in the Technical Appendices mirror those found in the full biennial report. The 2002 Technical Report and 2003 Technical Appendices document a variety of validity and reliability evidence. Many of the validity studies noted above are summarized in the Technical Report, along with evidence of test reliability. Note that the 2004 Technical Report will be available in July 2005.

8. **Other Studies.** The Abstracts of several additional, relevant studies are provided.

Impact on Getting to Proficiency:

The studies described in this staff note indicate that extensive resources and significant effort have been devoted to establishing both the reliability and validity of CATS. The goal is to establish, maintain and improve CATS to support and encourage efforts to improve the educational achievement of each child in Kentucky.

Groups Consulted and Brief Summary of Responses:

School Curriculum, Assessment and Accountability Council (SCAAC)
National Technical Panel on Assessment and Accountability (NTAPAA)

Both groups have supported CATS over the years. NTAPAA recently completed an official statement supporting the validity of the KCCT and CATS for the purposes for which they are intended.

Contact Person:

Dr. Bill Insko
Director, Assessment Implementation
Office of Assessment and Accountability
502-564-4394
binsko@kde.state.ky.us

Deputy Commissioner	**Commissioner of Education**

Date:

August 2005

ATTACHMENT A

Validation Studies and Evidence for Reliability

Validation Studies and Evidence for Reliability

According to KRS 158.6453 (5), "…the Department of Education shall gather information to establish the validity of the assessment and accountability program. It shall develop a biennial plan for validation studies that shall include, but not be limited to, the consistency of student results across multiple measures, the congruence of school scores with documented improvements to instructional practice and the school learning environment, and the potential for all scores to yield fair, consistent, and accurate student performance level and school accountability decisions. Validation activities shall take place in a timely manner and shall include a review of the accuracy of scores assigned to students and schools, as well as of the testing materials. The plan shall be submitted to the Commission by July 1 of the first year of each biennium. A summary of the findings shall be submitted to the Legislative Research Commission by September 1 of the second year of the biennium."

Scope of Work for Commonwealth Accountability Testing System Research

The complete Scope of Work for the Office of Assessment and Accountability, Kentucky Department of Education (KDE), was reviewed by the National Technical Advisory Panel and the Kentucky Board of Education, and was submitted to the Legislative Research Commission in August of 2000.

The research program described in the Scope of Work is undertaken to provide evidence as required by the statute cited above. The research activities discussed in this document cover a broad range including the following topics:

1. School Visit/Interview Study (p. 7)
2. Annual Multiple Assessments/Convergence Validity Evidence (p. 9)
3. Annual Third-Party Checking of KCCT Scaling and Equating (p. 11)
4. Annual Item Content, Item Difficulty, and Item Type Mapping for the Multiple KCCT Forms (p. 14)
5. Biennial School Classification Accuracy (p. 17)
6. Annual Student Classification Analysis (p. 20)
7. Reliability and Validity Technical Reports (p. 22)

Detailed Description of Each Study

A more detailed description of each study is presented in this section. Each study (except for topic number 7) is described in terms of:

- *Purpose (Why do the research?)*
- *Audience (Who will use the results of the research and how will they use it?)*
- *Methodology (How will the research be conducted?)*
- *Findings*
- *Recommendations*
- *Final Product*

1. School Visit/Interview Study

A Preliminary Examination of ACT Content and Instruction Alignment, HumRRO, March 2004

Purpose (Why do the research?)

Annual KDE validation-sponsored visits to Kentucky schools began in the 1996-97 school year for the purpose of linking the state's testing and accountability system to educational practices. Toward that end, the impact of the accountability system on teachers' instructional pedagogy, course content (or classroom curriculum), and teacher professional development have been examined and successful programs have been documented. KDE sponsored research has followed up on some issues and will continue to address additional topics that may arise. In addition, the visits/interviews provide a means to directly examine new topics that may arise. This provides an independent source of data about the positive, and sometimes, unexpected impact the CATS and the Kentucky Core Content Tests are having on districts, schools, and teachers.

The purpose of this 2004, preliminary study was to inquire about teachers' beliefs and practices vis-à-vis the ACT and the Kentucky Core Content Tests. The approach was to obtain examples of high school teachers' beliefs about the purposes of the KCCT (Kentucky Core Content Test) and ACT (American College Test) and their perceptions about how they prepare students for each of these tests. Results of the study are based on a convenience sample and are intended for preliminary examination. They are not to be generalized to the entire population of Kentucky teachers.

Audience (Who will use the results of the research and how will they use it?)

The results of the 2004 work will be informative to KDE leadership and the Kentucky Board of Education, groups interested in how Kentucky high school teachers might consider the demands of the ACT in planning and conducting their classroom instruction. The main purpose of the preliminary study, however, is to gather information for the design of a more extensive study.

Methodology (How will the research be conducted?)

In the 2004 study researchers interviewed 9 teachers -- 5 Kentucky public high school mathematics teachers and 4 English/Language Arts teachers, using a structured-protocol approach. This convenience sample of interviewees was selected via referral by Department of Education staff. Most were department heads at their schools and were known by the Department to be knowledgeable of the KCCT and the ACT. Interviews of approximately 40 minutes were conducted by telephone during February 2004. Data in the form of written notes were recorded by researchers – no voice recording was done. Notes were content analyzed, coded, and interpreted by the researchers.

Finding

- The spacing of KCCT tests may, in some cases, lead to the 'shifting' of the weight of instruction in a given content area to the tested grade, leaving little or no focus on that content in the preceding grade or grades. To prepare students for on-demand writing, eleventh- and twelfth-grade teachers may focus on writing to the near exclusion of reading. This could impact unfavorably students' performance on the ACT.
- Some teachers expressed uneasiness about their knowledge of the content of the ACT.
- Teachers may not be aware that American College Testing will field a writing assessment in February 2005.
- Teachers may focus their efforts on KCCT most of the year, turning to the ACT after the administration of the KCCT.
- Some teachers expressed the belief that some content in ACT is not included in Kentucky Core Content for Assessment, e.g., language mechanics, grammar, punctuation, rhetoric.
- Some teachers noted differences in the formatting and administration of ACT vs. KCCT, i.e., KCCT includes open-response items and is not timed.

Recommendations

No recommendations made, since this was a preliminary study. Results will be used to prepare a proposal for further study, if requested.

2. Annual Multiple Assessments/Convergent Validity Evidence

Purpose (Why do the research?)

Valid tests, by definition, produce test scores that behave in theoretically predictable ways. Therefore, an important method for establishing the validity of any given test is to systematically observe relationships between scores that the test produces and various other indicators that are expected to be associated with the test scores. Readily available data include student demographic data, CTBS/5 scores, student questionnaire data, and ACT data. These data allow examination of convergent and discriminant validity relationships as well as in-depth analyses of differences in performance related to gender, ethnicity, and SES. The student questionnaire data offer analyses of student characteristics (e.g. motivation, SES) in relation to test scores as well as students' perspectives on classroom instructional practices. The ACT data include students' self-reported transcripts, allowing analyses of course taking patterns and test performance. Any of these variables can be aggregated to the school level of analysis. Trends over time are also of interest.

Audience (Who will use the results of the research and how will they use it?)

All stakeholders with an interest in the valid interpretation of KCCT results, in the relationship of KCCT to nationally standardized tests, or in comparisons of KCCT disaggregated results to the disaggregated results of nationally standardized tests. Such individuals will find that these findings add to their understanding of Kentucky Core Content Test scores at both the student and school levels of analysis.

A Comparison of Students' KCCT and CTBS Scores Across Grade Levels, *HumRRO, 2004*

Methodology

This study investigated several correlational relationships between Kentucky students' KCCT and CTBS scores as they moved through the educational system, elementary to middle school, and middle school to high school. KCCT and CTBS results were analyzed separately. The investigation was longitudinal, i.e., it involved matching students' KCCT (or CTBS) test scores from one year to their scores on the same test the next time it was administered. Relationships between and within content areas were considered. Keep in mind that the CTBS is a nationally norm-referenced test measuring reading, language, and mathematics achievement at grades end-of-primary (third grade), sixth and ninth. CTBS results can be analyzed by content area or as a composite. Results of this study reflect the scores of all students taken together as well as disaggregated student groups -- gender, socioeconomic, and racial.

KCCT data were collected in the 2000 through 2003 administrations. Results of fourth-grade reading students in 2000, for example, were matched with the same students' results as seventh-grade readers in 2003. Similarly, scores of fifth-grade students who

took the KCCT mathematics test in 2000 were matched with the same students' eighth-grade mathematics scores in 2003. CTBS data used in the study were collected in the 2001 and 2004 administrations. Authors report that matching over school years was successful with respect to about 82% of the student scores in the data files.

Findings

- KCCT correlations over time were positive. Students who do well on the KCCT in the earlier grades, continue to do well in subsequent years.

- Correlational results in the same content area are strongest for KCCT mathematics, ranging from .68 (Grades 5 and 8, 2000 and 2003) and .74 (Grades 8 and 11, 2000 and 2003) (see pp. 5, 46 – 47).

- In some instances, within-cohort correlations (i.e., those produced by students in the same grade), *between* KCCT content areas (such as reading and science, both tested at fourth-grade), were higher than correlations in the *same* content areas, *between* grades (such as reading in 4th grade and reading in 7th grade). Grade 7 reading and science correlated .78 in 2000 (see report, p. 43). Grade 8 social studies and mathematics correlated .77 in that year (see report, p. 43).

- Students who do well on the CTBS in the earlier grades, continue to do well in subsequent years. Correlations within content areas range from .62 (Grades 3 and 6, Language, 2001 and 2004) to .73 (Grades 6 and 9, Mathematics, 2001 and 2004) (see report, p. 71).

- Within-cohort correlations, i.e., *between* CTBS content areas, but *within* grades and years, were as high as, or higher than, correlations between grades. For example, in 2004 the correlation between math and reading was .67, vs. a correlation of .66 for 3rd-to-6th grade math (2001 and 2004, p. 18) and a correlation of .64 for 3rd-to-6th grade Reading.

- Gender, racial, and SES differences remained relatively stable over time for both the KCCT and CTBS.

Relationships between Students' Scores on KCCT and CTBS, *HumRRO, 2004*

Methodology

While the paper discussed above examines student performance on the KCCT and CTBS separately, this paper examines them together, i.e., the relationship of students' performance on the KCCT at one grade level with the same students' performance on the CTBS at the next grade level. The authors conducted correlational analysis of test score data collected from 2000 through 2003. Their purpose was to investigate the range of the correlation coefficients produced on like vs. different subjects. In addition they compared correlation coefficients based on the performance of all students to those of disaggregated student groups. To perform the analyses the authors merged data files over

the two grades (for example, third and fourth), reporting that 83% of student cases were matched and included in the analyses.

Findings

Core content area correlations between like subjects ranged from r = .59 in reading (Grade 4, KCCT and Grade 3, CTBS, p. 53) to r = .74 in mathematics (Grade 9, CTBS and Grade 11 KCCT, p. 52). Core content area correlations between different subjects ranged from r = .53 (Grade 6, CTBS Mathematics and Grade 7 KCCT Reading, p. 48) to r = .63 (Grade 9, CTBS Mathematics and Grade 10, KCCT Reading, p. 49). Results shown by analyses of disaggregated group data are comparable to those of all students.

3. Annual Third-Party Checking of KCCT Scaling and Equating

Third-Party Checking of 2004 Scaling and Equating for the Kentucky Core Content Test, September, 2004

Third-Party Checking of 2003 Scaling and Equating for the Kentucky Core Content Test, September, 2003

Purpose (Why do the research?) 2003 and 2004

Psychometric scaling and forms equating (within and between years) of KCCT scores are conducted by two independent groups of contracted researchers. The researchers work in parallel, within the same operational timeframe, checking tables of intermediate results to ensure accuracy. Any quantitative discrepancy between the tables is identified, examined, and corrected. This step-by-step tracking procedure ensures the integrity of KCCT scores.

Audience (Who will use the results of the research and how will they use it?) 2003 and 2004

All stakeholders interested in the accuracy and integrity of KCCT scores will be assured by this work. The immediate audiences, however, are the primary contractor and the technical staff of the Office of Assessment and Accountability, Kentucky Department of Education. Other groups or individuals who wish to perform a technical review or to audit the Commonwealth Accountability Testing System will also appreciate these reports.

Methodology (How will the research be conducted?) 2004

With each annual iteration of the Kentucky Core Content Tests the primary contractor (currently CTB/McGraw Hill) and the third-party contractor (currently HumRRO) routinely duplicate all calibration sampling, scaling, and equating procedures using identical or comparable analytic procedures, programs, and technologies. The

methodology in 2004 was more involved than in preceding years, due to changes in reporting requirements provided by the 2001 reauthorization of the Elementary and Secondary Education Act, requiring states to report reading and mathematics results before schools convened for the 2004 – 2005 academic year.

The U.S. Department of Education agreed to a special arrangement in which Kentucky would report Preliminary Adequate Yearly Progress (AYP) results in reading and mathematics, based on multiple-choice items alone, in August of 2004. These results were to be followed by full results, based on multiple-choice and open-response items, in reading and mathematics in mid-October. Kentucky agreed that this arrangement would be in effect for 2004 reporting only. In 2004 special multiple-choice-only scales were created in mathematics and reading. The scales were developed in such a way as to link with the full 2002 multiple-choice and open-response scale. In 2005, full reading and mathematics results will be reported *prior* to the start of the school year as required in the federal statute while results of other content areas will be reported in September.

Findings in 2004

Sample Identification and File Construction

Four discrepancies arose and were resolved in the identification of 2004 calibration samples. First, the number of valid cases in the CTB and HumRRO calibration samples differed by one student. This student's record featured a blank in the CTB file and a zero in the HumRRO file. Second, CTB had initiated a new rule requiring that students produce at least one correct multiple-choice item response to be included in the calibration sample. It was found necessary to revert to the previous rule, i.e., students had to submit at least one response, not necessarily one that was found to be correct, to be included in the sample. Third, CTB's original files for each grade-by-subject group contained 50 – 70 fewer students than the HumRRO file, due to a programming error in which the CTB files were truncated, deleting the last few records. CTB repaired the error so that sample sizes used by both researchers matched. Fourth, a new procedure initiated in 2004, in which students marked answers directly in their test booklets, rather than on separate, scannable sheets, caused some difficulty. The original program created to process the multiple-choice responses made a false assumption about their configuration. This was corrected.

Scaling and Equating

An anomaly discovered in previous years, i.e., a difference between HumRRO and CTB in the final decimal place digit of some anchor-item parameters, was circumvented according to a previous agreement to use the parameters computed by HumRRO via its SAS program (which rounds to the final decimal place) versus the CTB program (which truncates). It was also discovered that the Stocking-Lord constants M1 and M2 used to linearly map the 2004 z-score metric to the 2002 scale-score metric were slightly different in grade 10 reading. Investigation showed that CTB had inadvertently used the smaller sample (less 50 – 70 students as described above) in calibrating. This was corrected.

HumRRO and CTB reached near-exact agreement on the equating constants for all grade-by-subjects in reading and mathematics in all content areas (multiple-choice-only and full-set analysis). HumRRO also verified the cut-points on the raw-score-to-scale-score tables. In one case HumRRO and CTB assigned different performance categories to a scale score, even though both groups of researchers had identical data for the student. According to previously agreed-upon rules, students are assigned to the Novice-Non-performance category when their raw scores are at chance level. In this case the raw score was just above chance. CTB agreed to reassign the scale score to Novice–Medium, reconciling the data files.

Methodology (How will the research be conducted?) 2003

The primary contractor (currently CTB/McGraw Hill) and the third-party contractor (currently HumRRO) duplicated all sampling, scaling, and equating procedures using identical or comparable analytic procedures, programs, and technologies. Discrepancies were shared among the analysts, investigated, and resolved.

Findings in 2003

In 2003 Kentucky's primary contractor, CTB-McGraw Hill, conducted all scoring, rather than subcontracting parts of the scoring, as in previous years. As a result, HumRRO, the third-party contractor, was obliged to alter its data processing protocols and statistical programs to render the data sets compatible with CTB proprietary analysis software (Pardux and Flux). At the same time, the two contractors used different statistical programs in constructing their calibration samples. In the course of tracking intermediate results, they discovered unequal numbers of students in their respective calibration samples, the largest difference being small in comparison to sample sizes of 40,000 or more students. The contractors resolved this discrepancy by agreeing to use the sample with the smaller number of students.

KCCT forms are equated within and between years using anchor items (i.e., items that remain in the test across forms and years.) An issue arose in the context of anchor items used in fourth-grade science. The HumRRO anchor file was found not to match the CTB anchor file, each having been created using different programs. It was discovered that the discrepancy was the result of incorrect item numbering and was corrected. A discrepancy in the last decimal point of some item parameters was discovered to be the result of a difference in the programs used by the two contractors, i.e., the truncation procedure used by the CTB program (FLUX), versus a rounding procedure used by the HumRRO SAS program. This last issue, discovered and investigated in 2002, was found to result in very minor final differences, causing no discrepancies in student or school classification. The raw-score-to-scale-score tables created by HumRRO and by CTB did, however, show minor differences, no larger than one scale score point. The differences did not affect student performance classification.

4. Annual Item Content, Item Difficulty, and Item Type Mapping for the Multiple KCCT Forms

Item Content and Difficulty Mapping by Form and Item Type for the 2004 Kentucky Core Content Tests*, HumRRO, 2004.*

Item Content and Difficulty Mapping by Form and Item Type for the 2003 Kentucky Core Content Tests, *HumRRO, 2003.*

Purpose (Why do the research?)

The valid interpretation of test scores used for school accountability requires strong content-related validity. This in turn requires evidence of the comparability of test forms with respect to content and difficulty. One line of evidence for comparability may be brought forward in the course of answering the following questions:

1. Is the Kentucky Core Content for Assessment equitably represented on each form?
2. Is the Kentucky Core Content for Assessment proportionally represented by Novice, Apprentice, Proficient, and Distinguished (NAPD) proficiency categories?
3. How is the Kentucky Core Content for Assessment represented by item format (multiple choice versus open response)?
4. What is the distribution of item formats by Novice, Apprentice, Proficient, and Distinguished (NAPD)?

Audience (Who will use the results of the research and how will they use it?)

This information will facilitate judgment of forms comparability by KDE, KBE, the National Technical Advisory Panel for Assessment and Accountability, and other stakeholders. The test-construction contractor will gain feedback for use in subsequent test construction.

Methodology (How will the research be conducted?)

Item mapping is a graphical approach to addressing the questions posed above. It involves plotting variables that reflect item content, item difficulty (N, A, P, D), and item format (multiple-choice vs. open-response) on coordinate axes. For example, item difficulty (in terms of N, A, P, D) may be plotted against test form (numbered 1, 2, 3, 4, 5, and 6) to clarify the representation of items of each difficulty level on each form.

The Kentucky Core Content for Assessment (KCCA) and Program of Studies is formally organized by subdomain and by section within subdomain. Achievement levels (or proficiency categories) were set on the scoring scale in 2000 and 2001, using judgment-eliciting procedures in which input from hundreds of Kentucky teachers was obtained. The process resulted in scores corresponding to cuts (or divisions) on the continuum of scores ranging from 325 to 800. The cut-scores were chosen to best distinguish student

performances described by the four achievement levels -- Novice, Apprentice, Proficient, and Distinguished. Since all items on the KCCT are calibrated on this scale, it is possible to classify all items according to achievement level.

For each content domain test at each tested grade, the authors created two-way graphs, formatted as stacked bar charts. The graphs illustrate relationships between balance in coverage, proportionality, and comparability of test items. While these goals are achieved through careful test construction, the graphs illustrate the extent to which balance, proportionality, and comparability have been achieved. Graphs were also constructed to address another matter of interest, i.e., how achievement level is associated with one item format. The table below shows how test construction goals displayed on item maps are associated with issues that must be monitored.

Test Construction Goal	Item Map/Graph Dimensions	To Be Monitored …
1. **Balance in Coverage** of Subdomains and Subdomain Sections by Forms	Subdomain/section by KCCT form	Subdomain items tend to be present on some forms more than others
2. **Proportionality** Between Achievement Level and KCCA Content Subdomain	Subdomain/section by difficulty/achievement level (N, A, P, D).	Items in some subdomains may be easier (or more difficult) than items in other subdomains.
3. **Balance in Coverage** of Subdomain and Subdomain Section by MC vs OR Item Formats	Subdomain/section by item type (MC or OR)	Item formats not be used in equal proportions in all subdomains
4. **Comparability of Forms** in Difficulty/ Achievement Level	Difficulty/achievement level by KCCT form.	Non-comparability -- Some forms more difficult or easier than others

2003 and 2004 Findings

The table below lists the 2003 and 2004 Kentucky Core Content Tests for which issues of balance, proportionality, and comparability were noted. Notice that issues identified in 2003 were, with one exception, not identified in 2004.

Content Tests by Mapping Graph		
Mapping Graph	**2003**	**2004**
1 Subdomain/Section by Form	Reading: Grades 4 & 7	Reading: Grades 4 & 10
2 Subdomain/Section by Difficulty	Science: Grade 4 & 11 Math: Grade 8 Soc. Studies: Grade 5 Arts & Hum: Grade 8	
3 Subdomain/Section by Item Format	Science: Grade 4 Math: Grade 8 Social Studies: Grade 5 Arts and Hum: Grade8	
4 Difficulty by Form	Reasonable balance for each grade/content .	
5 Item Format by Difficulty	Arts and Humanities: Grade 5	

This report supports the general conclusion that the distribution of items at each subdomain/section by form tends to be balanced (Goal 1); the distribution of items at each subdomain/section by difficulty level tends to be proportional (Goal 2); the distribution of subdomain/section items by item format tends to be balanced (Goal 3); the difficulty of items is comparable for each form. As might be expected, the graphs illustrate that there is an association between item format and achievement level.

Recommendations

The report suggests the need for test-construction contractor and Content Advisory Committees (CACs) to consider the use of item format by Core Content to ensure that any imbalance or disproportion is content driven. In addition, CACs may need to consider increasing the difficulty range of multiple-choice items used in most KCCT assessments.

5. Biennial School Classification Accuracy

The Accuracy of School Classifications for the 2004 Biennium of the Kentucky Commonwealth Accountability Testing System, Hoffman and Dickinson, May 2005.

Purpose (Why do the research?)

Biennial accountability classification of schools into the three school categories (Meets Goal, Progressing, or Assistance) is based on the *relationship* between a school's biennial accountability index (a composite of test scores and indicators) and its biennial accountability goal. Schools that meet their goals are classified as Meets Goal, while those that fail to meet their goals are classified either as Progressing or in Assistance. Schools that exceed their Assistance point, but fail to reach their Biennial Goal are classified as Progressing, while those whose indices are equal to or below the Assistance point are classified as in Assistance. This report provides evidence in support of the claim that Kentucky school classification accuracy is acceptable and appropriate.

Audience (Who will use the results of the research and how will they use it?)

Accuracy of school classification is essential to the credibility of the state accountability program. It is therefore of interest to educators and parents as well as policy makers, such as the School Curriculum, Assessment and Accountability Council (SCAAC), and the National Technical Advisory Panel on Assessment and Accountability (NTAPAA).

Methodology (How will the research be conducted?)

To estimate classification accuracy the authors applied Bayes Theorem to estimates of the distribution of possible true scores associated with observed scores. These estimated distributions are presented below, in *Findings*, Tables 1 and 2. (Baye's Theorem demonstrates how to compute conditional probabilities.) To perform the accuracy analyses, estimates of standard errors were required. Using generalizability analysis, the authors computed standard errors for each of the 27 CATS tests (reading, writing on-demand, writing portfolio, mathematics, science, social studies, arts and humanities, practical living/vocational studies, and NRT) by school-level combinations, for schools of small, medium, and large sizes. This resulted in 81 separate standard error estimates for the KCCT. This is because the standard error of the Biennial Accountability Index, a composite, is based on the standard errors of all its elements.

Findings and Recommendations

Kentucky's biennial accountability classification of schools as Meets Goal, Progressing, or Assistance is based on the *relationship* between a school's biennial accountability index (a composite of test scores and non-academic indicators) and the school's biennial accountability goal and assistance point. Due diligence requires the Department of Education to provide evidence of the accuracy of school classification. The findings of this paper support the claim that Kentucky school classification accuracy is acceptable and appropriate.

To minimize the likelihood that a school may be incorrectly assigned to a category below its true performance level, school biennial goals are adjusted by the size of the standard error; school accountability indices are then compared to this figure for accountability classification.

The analysis utilizes the Classical Measurement Theory concept of *true score*. A true score is one without error. The final outcome of the present classification analysis is an estimate of the probability that, for each school, the "true" (but unknowable) classification is the same as the classification actually obtained. As reflected below in the percentages printed on the diagonal of Table 1, in 2004, 82% of schools were accurately classified when the standard error, or fairness margin, was taken into account. The 82% is the sum of 45% of schools Meeting Goal, 34% Progressing, and 3% in Assistance.

Table 1
Classification Probabilities for 2004 School Accountability

Expected True Category	Assigned Category (Before Novice and Drop Criteria Applied)			Total Expected for True Classifications
	Meeting Goal	Progressing	Assistance	
Meeting Goal	***45%***	*1%*	*0%*	46%
Progressing	*11%*	***34%***	*1%*	46%
Assistance	*2%*	*3%*	***3%***	8%
% in Observed Class	58%	38%	4%	100%
Number in Obs Class	694	461	47	1202

Note: Bold italics numbers indicate expected probabilities of accurate classifications. They sum to 82%. Only schools with data for all four years and with constant grade configurations are include in the analysis.

Table 2, below, shows how accurate the accountability system would be if schools were classified without the baseline SEM safety net. These results are a better indication of measurement accuracy. Without the safety net, schools would be assigned to the category most likely to contain their true scores. The baseline safety net increases the total number of schools that are classified as Meets Goal, thereby reducing the probability of unjustifiably classifying schools below their true category. The result is that some percentage of schools are over-classified. When the standard error is ignored, 89% of schools are accurately classified. See Table 2, below.

Table 2
Classification Probabilites for 2004 School Accountability without Baseline Safety Net

Expected True Category	Observed Category Without Applying Baseline SEM Safety Net			Total Expected for True Classifications
	Meeting Goal	Progressing	Assistance	
Meeting goal	***44%***	*2%*	*0%*	46%
Progressing	*5%*	***38%***	*3%*	46%
Assistance	*0%*	*1%*	***7%***	8%
% in Observed Class	49%	41%	10%	100%
Number in Obs Class	586	499	117	1202

Note: Bold italics numbers indicate expected probabilities of accurate classifications. They sum to 89%. Only schools with data for all four years and with constant grade configurations are include in the analysis.

To further clarify any possible misclassification and to assist in achieving the highest possible performance, schools that are classified as needing Assistance are provided the opportunity to participate in a Scholastic Audit or Review.

6. Annual Student Classification Analysis

The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the Kentucky Core Content Test, HumRRO, 2003 and 2004.

Purpose (Why do the research?)

The Kentucky Core Content Tests are administered annually to students in 18 different grade-by-subject combinations (e.g., Grade 4 Reading; Grade 8 Mathematics). Based on their scale score, students are classified into one of four performance categories (Novice, Apprentice, Proficient, and Distinguished, often referred to as NAPD). Results are used to compute school accountability index scores.[1] It is important to examine the accuracy of these student score classification decisions to be fair to students and schools.

Audience (Who will use the results of the research and how will they use it?)

Estimates of the accuracy of student performance scores are helpful to schools, students, parents, educators, policy makers (including the School Curriculum, Assessment and Accountability Council, or SCAAC), technical reviewers (including the National Technical Advisory Panel on Assessment and Accountability, or NTAPAA), and others. This knowledge provides a basis for understanding the strength of classification, while at the same time illustrating the need to avoid inappropriate use of student performance scores (e.g., making instructional decisions about individual students on the basis of KCCT performance scores, or indeed, any test score, alone, independent of other relevant information).

Methodology (How will the research be conducted?)

Student classification accuracy was examined using a method developed by HumRRO and approved by NTAPPA in 1999. Readers interested in a detailed description of the methodology are referred to Hoffman, R. and Wise. L, 1999, *Establishing the Reliability of Student Level Classifications: Analytic Plan and Demonstration*. A summary of the approach follows.

Classification error occurs, for example, when a student who truly performs at the Proficient level is scored as Distinguished, or when a student who is truly Apprentice, is scored as a Novice. Recall that assignment of scores to performance level categories is done using scale scores (not raw scores). IRT analysis transforms raw test score (number correct) to a scale-score metric (such as the KCCT scale of 325 – 800). Agreed-upon

[1] Novice and Apprentice categories for selected subjects are each divided into three levels: low, medium, and high. Accuracy of assignment to these subcategories was not included in the analysis.

cut-points are then used to separate the scale-score continuum into four intervals corresponding to N, A, P, and D performance levels. Student test scores are classified as N, A, P, or D, depending upon where they lie on the 325 – 800 scale-score continuum.

According to Classical Test Theory, any given test score is a composite of the true score and measurement error. KCCT scale scores, therefore, include some measurement error. While the raw-score estimate of measurement error is constant across possible scores on a given test, the error estimate associated with scale scores varies with the measurement scale. The beginning and end points of the scale are associated with larger error size than scores in the middle. These IRT-generated error estimates are called *conditional* standard errors of measurement. Estimates of classification accuracy must take the conditional standard error of measurement into account.

The authors suggest that, again, based on Classical Test Theory, for every observed test scale score, there is not just one true score, but a *distribution* of possible true scale scores. The possible true scores are located around the observed score. One can estimate the probability of each true score in the distribution around the observed score, using the conditional standard error of measurement provided by IRT analysis. Then, applying Bayes' theorem, one can compute the probabilities of correct and incorrect scale-score classifications.

The authors report the probability of accurate vs. inaccurate classification in each performance category in four-by-four tables such as the one provided below. The sum of the probabilities in the 16 cells of each table is 1.00. The probability of accurate classification in each category is printed in bold. By adding the probabilities appearing in the diagonal cells, one can determine the total percentage of expected correct assignments. In Table A-5, below (Hoffman, 2004, p. 10), Grade 8 Mathematics 2004, the sum of the probabilities on the diagonal is 80.46%. The authors present 18 classification probability tables, one for each school-level by core content combination.

Table A-5. Grade 8 Mathematics 2004
Percentages of True Scores Being in Assigned Classification

	Assigned Classification				Total %
True Class	Novice	Apprentice	Proficient	Disting.	Expected
Novice	***21.87***	*3.61*	*0.00*	*0.00*	25.49
Apprentice	*3.95*	***33.81***	*4.36*	*0.00*	42.12
Proficient	*0.00*	*3.62*	***18.99***	*2.25*	24.86
Distinguished	*0.00*	*0.01*	*1.74*	***5.79***	7.54
Total % Assigned	25.82	41.05	25.10	8.03	100.00
Total % Expected Correct Assignments: **80.46**			Average Distribution Error: 0.54		

The authors also point out a shortcoming of this methodology, namely, that it assumes the random distribution of all types of measurement error (related to students, test items,

etc.) across test forms. They suggest that a better approach to accuracy analysis would include an empirical study of the level of error due to forms, based on a random sample of students who are asked to respond to two different test forms on different occasions (p. 18). The size of this type of error could then be considered in classification accuracy analysis.

The table below presents the total percentage of correct assignments for each school level and core content area reported by the authors for tests administered in 2003 and 2004.

Total Student-Level Classification Accuracy by Core Content Test, School Level, and Academic Year

	Elementary		Middle		High	
Core Content Test	**'02 – '03**	**'03 – 04**	**'02 – '03**	**'03 – '04**	**'02 – '03**	**'03 –' 04**
Reading	78.71	80.70	79.90	81.47	81.14	82.99
Mathematics	74.29	72.68	80.21	80.46	79.60	78.57
Science	76.62	74.80	73.07	71.80	77.16	76.80
Social Studies	71.41	69.76	79.51	79.31	80.11	78.92
Arts & Humanities*	62.16	67.56	60.29	67.12	65.29	63.93
PL/VS*	58.77	56.77	61.78	63.65	61.44	63.07
*Because Arts & Humanities and PL/VS are shorter tests, the classification accuracy is expected to be lower in comparison to the other content areas.						

7. Technical Reports (Reliability and Validity Evidence)

Reliability Evidence

The Commonwealth Accountability Testing System, a high-stakes, standards-based accountability system, sets consequences for schools on the basis of their performance on state tests and other non-academic indicators. By statute the Kentucky Department of Education is required to ensure that its performance measurement instruments produce valid and reliable scores. To ensure reliable measurement, sources of measurement (and sampling) error must be identified, steps must be taken to minimize the size of error, and error estimates must be included in evaluation of results. The primary sources of measurement error in standards-based achievement testing involve item development and test construction, scoring, and the classification of students and schools. The latter was addressed in sections 5 and 6. Other sources of error are briefly discussed below.

<u>Item Development and Test Construction</u>

Test items and prompts must be written in such a way as to generally elicit correct responses from students who know and understand the relevant content. This is likely to occur when the language of test items is grammatical, clear, and grade-level appropriate.

At the same time, students who do not know and understand the relevant content should not generally respond correctly to test items. Although some measurement error is likely to occur, even when items are properly constructed, if the error level is known, it can be taken into account when test scores are reported and interpreted. The reliability of a group of test items (on a test form, for example) is often statistically estimated using procedures such as coefficient alpha.

Test forms within a content area must be comparable with respect to the difficulty of items, both within, and between, years of test administration. Although the content of the forms of the KCCT varies by design, allowing full coverage of the Kentucky Core Content each year at the school level, an attempt is made, when constructing the forms, to balance them in terms of difficulty. Comparability of forms requires meticulous form construction (selection of items) and psychometric equating within and between years. The comparability of forms can be estimated psychometrically using item analysis and test characteristic curves (an Item Response Theory technique). It can be demonstrated visually using item-difficulty maps (discussed above, Section 4).

Scoring

Scoring of student responses to multiple-choice items is done electronically. Although minimal, errors may result from improper coding and extraneous marks on scannable response sheets. The size of this sort of error is thought to be minimal, due to the cautionary advice given to test administrators. Open-response items, on the other hand, are vulnerable to scoring error due to differences in raters. To minimize such error (and maximize reliability) the test contractor responsible for scoring takes due care in training raters and in monitoring the scoring process, and recording estimated levels of inter-rater agreement.

The Kentucky Department of Education advises against making student-level evaluations or decisions based on any one measurement, such as a KCCT score, without taking additional supporting information into account. Teachers and school administrators are advised to consider any test score in the context of other student work, including other test scores, as well as teachers' evaluations of experiences with the student, even when test reliability is very high. Readers should keep in mind, however, that KCCT test reliabilities compare very favorably with reliabilities reported by publishers of nationally-normed, standardized achievement tests. For example, KCCT reliabilities are comparable to those reported by ACT and CTBS.

Coefficient alpha is a statistic used to estimate the consistency of student responses, including responses to open-response items, over many test items. The table below presents KCCT student-level coefficient alpha statistics for 2000 through 2004 by grade and core content area. Median alpha values and their ranges are computed across the 6 forms in six subjects using both multiple-choice and open-response items and across 12 forms in Arts & Humanities and Practical Living/Vocational Studies. All values are based on data contributed by students who were eligible to complete testing and who were present on the day of testing. The responses of absent students (whose test booklets have all blanks) are excluded to avoid overestimation of score reliability. When a test

booklet includes at least one response, zeros corresponding to any blank items are entered into the data file and subsequently included in the computation of coefficient alpha.

KCCT Reliability 2000 – 2004
Coefficient Alpha Median and Range by Subject by Grade

Subject by Grade		2000		2001		2002		2003		2004	
		Median[1]	Range	Median[1]	Range	Median[1]	Range	Median[1]	Range	Median[1]	Range
4/5	**Reading**	.88	.87-.88	.88	.87-.88	.88	.86-.88	.86	.85-.87	.87	.84-.87
	Mathematics	.87	.86-.88	.87	.86-.88	.87	.86-.88	.87	.86-.87	.86	.83-.88
	Science	.84	.80-.85	.84	.80-.85	.83	.81-.84	.83	.82-.85	.84	.81-.85
	Social Studies	.84	.84-.85	.84	.84-.85	.85	.83-.86	.84	.83-.85	.83	.82-.84
	A & H	.66	.59-.70	.66	.63-.67	.66	.63-.71	.66	.62-68	.64	.56-.73
	P/VS	.63	.51-.65	.63	.53-.67	.69	.67-.73	.61	.50-.64	.58	.49-.67
7/8	**Reading**	.87	.87-.88	.87	.87-.88	.87	.87-.88	.86	.85-.87	.86	.85-.87
	Math	.89	.88-.90	.89	.88-.90	.89	.88-.90	.89	.88-.89	.89	.88-.90
	Science	.84	.83-.86	.84	.83-.86	.86	.84-.86	.85	.84-.86	.84	.84-.86
	Social Studies	.89	.87-.89	.89	.87-.89	.88	.87-.89	.88	.87-.89	.88	.86-.88
	A&H	.68	.58-.70	.70	.66-.73	.69	.67-.73	.67	.59-.73	.66	.61-.73
	PL/VS	.69	.63-.72	.70	.66-.74	.71	.67-.74	.68	.63-.73	.66	.62-.71
10/11[2]	**Reading**	.88	.87-.89	.88	.87-.89	.88	.88-.89	.87	.87-.88	.89	.88-.91
	Mathematics	.88	.85-.89	.88	.85-.89	.89	.87-.89	.89	.88-.89	.89	.88-.90
	Science	.84	.82-.85	.84	.82-.85	.85	.81-.86	.84	.82-.85	.83	.82-.84
	Social Studies	.88	.87-.88	.88	.87-.88	.89	.88-.89	.88	.87-.89	.88	.87-.89
	A&H	.66	.58-.69	.67	.61-.72	.69	.65-.72	.66	.62-.68	.66	.57-.71
	PL/VS	.65	.62-.67	.64	.60-.68	.65	.62-.68	.64	.56-.67	.63	.57-.71

[1]Median coefficient alpha based upon operational matrix open-response and multiple-choice items across the 6 or 12 forms of the KCCT.
[2] Reading and PL/VS are tested in the 10th grade.

Since coefficient alpha estimates reliability over many test items, it is not possible to use it to measure reliability of the Writing Portfolio, each submission of which receives only one holistic score. However, we do have evidence supporting the reliable scoring of writing portfolios. In 2004, 101 schools were identified for writing portfolio audits. Seventy-five schools were selected at random while the remainder were purposefully selected. For the Random group 75.10% of the locally assigned scores were confirmed and for the Purposeful group, 76.59% of the portfolio scores were confirmed by one or more audit readers.[2] For both Random and Purposeful groups, over 99 percent of re-scoring were within an adjacent category of the original teacher holistic writing portfolio score.

[2] CTB McGraw-Hill (2004) Kentucky Writing Portfolio Audit Final Report, p. 36.

Validity

The Kentucky Department of Education intends the results of the Kentucky Core Content Tests to be used for three purposes:

1. Accountability – state and federal evaluation of school performance
    - Kentucky bases accountability on school accomplishment of target goals set on the Biennial Accountability Index, an arithmetic composite of student test scores in the seven tested content areas and academic indicators such as attendance.
    - The United States Department of Education bases accountability on school achievement of AMOs (Annual Measurable Objectivities) in reading/language arts and mathematics. AMOs are set in terms of percentages of students scoring Proficient or above on the Kentucky assessments.
2. Improvement of instructional programs (in core content areas) directed toward all students in grade-, school- or district- level groups as well as toward specific disaggregated student groups who attend a given school (or school district).
3. General feedback to parents on the academic performance of individual students.

The Kentucky Department of Education contends that school scores based on the results of the Kentucky Core Content Tests accurately reflect the performance of a school's students as a whole, as well as students in school-level, subject-by-grade populations, when used for accountability, school improvement, and general reporting to parents. In support of this claim it submits a body of evidence accrued over time through the Department's research and development efforts guided by the National Technical Advisory Panel on Assessment and Accountability.

Content-Related Validity Evidence

Baker and Linn (2002, p. 7) suggest that evidence of the content-related validity of assessments must address two central questions:

1. Is the definition of the content domain to be assessed adequate and appropriate?

2. Does the test provide an adequate representation of the content domain the test is intended to measure?

The following discussion addresses these central questions.

Appropriate Definition. The Kentucky Core Content for Assessment and Program of Studies serve as the definition of the content domains in which Kentucky tests its students. It is argued that the domains are appropriately defined, because, they reflect twelve years of guided effort on the part of experienced Kentucky educators and content-area specialists.

Adequacy of Representation. It is argued that the subject domains addressed by the Kentucky Core Content for assessment are adequately represented on the Kentucky Core Content Tests by virtue of a responsible test construction process guided by the test blue print. Readers are referred to the 2002 Technical Report and the 2003 Technical Report Appendices. Test items and scoring rubrics are developed by experienced Kentucky teachers who are recruited to serve on Content Advisory Committees (CACs). Contracted test development professionals edit items provided by the teachers and construct the test forms, in accordance with the test blueprint, mapping categories of the core content (and associated intellectual processes) to specific test items. Test items are field tested and evaluated psychometrically prior to their operational use in measuring student performance.

The full test blueprint, tables of test specifications mapping categories of the core content (and associated intellectual processes) to specific test items, provides more concrete evidence. These tables may be found in the 2002 Technical Report along with the academic expectations. Referencing Standard 13.3 of *Standards for Educational and Psychological Testing* (1999) Baker and Linn indicate that

> Detailed analyses of the relationship between the content domain of the content standards and the specific content of the assessment are needed to support such inferences. Confirmation of alignment of the test items and content standards by independent judges provides one type of evidence. This may be accomplished by having judges assign assessment tasks to the content standards they believe the tasks measure and comparing those assignments with the assignments of the developers of the assessment tasks.

Concurrent Evidence

Achievement Tests. The validity of test use and interpretation may be supported by indications of the close (or distant) relationships among scores on tests designed to measure similar constructs. However, state assessments such as the KCCT are generally thought of as content-standards-referenced tests in the jurisdictions that administer them. This is to say that the tests are unique -- there are no comparison tests measuring all of the knowledge and skill content as well as the intellectual processes provided in the content and performance standards. However, comparisons between nationally norm-referenced assessments and the KCCT in the same general content areas are reasonable in view of the fact that such tests have some overlap with the KCCT. Keep in mind, however, that the KCCT includes open-response items that require students to use intellectual processes that may not be elicited by multiple-choice test items unless specifically designed to do so.

Consider that in Kentucky few achievement tests other than the KCCT are administered to all students, or even to a representative sample of students, in a given school. Although many eleventh-grade students who seek higher education take the ACT, this subpopulation of students does not reflect the total population of Kentucky students. Examination of the relationship between student performance on the KCCT and that on

the ACT must take this fact into account. Bacci et al.[3] confirm previous results shown by Hoffman with respect to KIRIS. Using only the ACT-taking population of students at a school, he and colleagues found that schools with high scores on the ACT also have high open-response scores on KIRIS. At the school level of analysis, GPA is not related to either students' open-response performance or to performance on the ACT. This is presumably due to differences among schools in their grading standards. When gains in schools means are calculated using only the ACT-taking population of students, schools whose ACT-taking students are gaining on any one of the assessments tend to gain on all of the assessments, including open-response, ACT, and GPA. This result is obtained in spite of the typically unstable nature of correlational examinations of score gains.

The National Assessment of Educational Progress. The National Assessment of Educational Progress (NAEP) is a congressionally authorized, standards-referenced, national student achievement assessment. It is administered in reading and mathematics to a randomly-selected, representative sample of 4th and 8th grade students in each participating state every two years. Results of successive years of administration may be compared over time and NAEP state results can be compared to the nation as a whole or to those of other states. NCLB, which requires the participation of every state in the state NAEP test, also provides that state NAEP results be used to evaluate the quality of state accountability programs. The logic is that, if improvement in student learning is real, rather than artifactual, improvement should be demonstrable on NAEP as well as on state-required assessments. While the statute does not refer to NAEP as the criterion test for state assessments, the implication is that states must seriously consider their NAEP results in the context of their state accountability results.

Kentucky NAEP results are available in reading and mathematics for both 4th and 8th grade for the 2000 and 2003 administrations. Prior to 2000, mathematics and reading NAEP tests were administered in alternate years. The mathematics test was administered in 1996, the reading test, in 1998. In compliance with a new biannual schedule, both are to be administered in spring 2005 and results are expected in late summer of 2005. This will provide three years of parallel results. The following presents highlights of Kentucky 2003 results.

- **NAEP Fourth-Grade Reading**

    - The percentage of Kentucky fourth-graders scoring at or above Proficient in 2003 was 31%, comparable to (not significantly different from) the percentage of students in the nation scoring at this level in 2003 -- 30%. Sixteen states scored at the national average, while 13 states scored significantly below the nation, and 24 states scored above the national average. (Source: *The Nations Report Card, Reading Highlights,* 2003)

[3] Bacci, E. D. (2003) Relationships Among Kentucky's Core Content Test, ACT Scores, and Students' self-Reported High School Grades. Radcliff, KY: Human Resources Research Organization.

- Kentucky fourth-grade readers' average scale score of 218 in 1998 exceeded the national average of 213. The Kentucky score increased slightly to 219 in 2003, exceeding the new national average of 216.

- The percentage of Kentucky fourth-grade readers scoring at Basic or above in 1998 was 62%. This increased to 64% in 2003. At the same time the percentage scoring at or above Proficient increased from 29% in 1998 to 31% in 2003.

- **NAEP Eighth-Grade Reading**

  - Thirty-four percent of Kentucky eighth-grade reading students scored at or above Proficient in 2003. Their achievement was comparable to (not statistically different from) the national average at 30% Proficient or above. Seventeen states scored below the nation and 25 states scored above the nation. (Source: *The Nation's Report Card*, *Reading Highlights,* 2004)

  - Kentucky's eighth-grade readers, slightly exceeding the national average scale score of 261 in 1998 with score of 262, improved to a score of 266 in 2003, again exceeding the new national average scale score of 261.

  - The percentage of eighth-grade readers scoring at or above Basic increased from 74% in 1998 to 78% in 2003, while the percentage of readers scoring at Proficient or above increased from 30% to 34% in the same interval.

- **NAEP Writing**

  Below are results of NAEP writing assessed at the 4^{th} and 8^{th} grade level in 2002, and at the fourth-grade level in 1998 (8th grade level not assessed in 1998).

- **NAEP Fourth-Grade Writing**

  - In 2002 27% of Kentucky fourth graders scored Proficient or above, the same as the national average of 27%. Kentucky's writing score was higher than that of the 26 states and jurisdictions whose performance was significantly below the national average, and 10 jurisdictions performed significantly above the national average. Kentucky's 2002 fourth-grade average writing scale score was 154, vs. the national average of 153.

  - While 27% of Kentucky fourth graders scored at Proficient or above in writing, 59% scored at Basic and 14% fell below the Basic level.

- **NAEP Eighth-Grade Writing**

  - At 25% Proficient or above in 2002, Kentucky was among 22 states whose eighth-grade writing students scored significantly below the national average of 30% Proficient or above. Fifteen jurisdictions scored at the national average and ten jurisdictions scored significantly above the average.

  - Kentucky's 2002 eighth-grade' writing scale score was 149; although lower than the national average scale score of 152, it exceeded the 1998 score of 146.

  - While 25% of 2002 eighth-graders scored Proficient or above, 60% scored at Basic, and 15% scored below the Basic level.

- **NAEP Mathematics**

The following discusses the results of assessments administered in 2000 and 2003. Mathematics will be administered in spring 2005.

- **NAEP Fourth-Grade Mathematics**

  - Kentucky scored among the 16 states and jurisdictions whose fourth-grade mathematics performance in 2003 was significantly lower than that of the nation. Twenty-two percent of Kentucky students scored at Proficient or above, while 32% was the national average. The achievement of 19 states was comparable to that of the nation as a whole and 18 states scored significantly higher than the nation (*Mathematics Highlights*, 2003).

  - Kentucky's fourth graders went from a mathematics scale score of 219 in 2000 to a score of 229 in 2003. The national average mathematics scale score of 224 in 2000 increased to 234 in 2003.

  - Scale-score gains in fourth-grade student math performance in 2003 are reflected in increases in the percentage of fourth-grade students scoring at Basic or Above – 72% in 2003 vs. 59% in 2000, and the percentage Proficient or above, 22% in 2003, vs. 17% in 2000.

- **NAEP Eighth-Grade Mathematics**

  - Twenty-four percent of Kentucky's eighth graders scored at Proficient or above in 2003 mathematics, significantly below the national average of 27%; 17 states and jurisdictions performed below the national average, 12 performed at average, and twenty-four states performed above the national average percentage (*The Nation's Report Card, Mathematics Highlights, 2003*).

- With a an average mathematics scale score of 270 in 2000, Kentucky's eighth graders missed the national average scale score of 272, and did so again in 2003 with a state average of to 274 , vs. the national average of 276 in 2003.

- Scale-score improvements in eighth-grade mathematics performance are accompanied by improvements in the percentage of students scoring at Basic or above, 65% in 2003 vs. 60% in 2000, and increases in Proficient or above -- 24% in 2003, vs. 20% in 2000.

- **NAEP Science**

  The following discusses the results of NAEP science, assessed at the 4th- and 8th-grade levels in 2000 and at 8th-grade in 1996 (science not assessed at 4th-grade level in 1996).

- **Fourth-Grade Science**

  - Twenty-nine percent of Kentucky's fourth-graders scored Proficient or above in science in 2000, comparable to (not significantly different from) the national average of 27%. The scores of 16 other states were comparable to the national average, while 15 states scored significantly below, and 12 states significantly above, the national average.

  - In 2000, the average scale score of Kentucky's fourth-grade science students was 152, as compared to the national average of 148. While 29% of Kentucky's students scored at Proficient or above, 41% of fourth-graders scored at Basic, and 30% scored below Basic in 2000 science.

- **Eighth-Grade Science**

  - Twenty-nine percent of Kentucky eighth-grade science students scored at or above Proficient in 2000, comparable to (not significantly different from) the national average of 30% Proficient or above. Seven other states performed comparably, while 17 jurisdictions significantly exceeded, and 17 scored significantly lower than, the national average.

  - In 2000, Kentucky's eighth-grade science students' average scale score was 152, as compared to the national average of 149. In 1996 Kentucky's average eighth-grade scale score was 147, as compared to the national average of 148. While 29% of eighth-grade science students scored Proficient or above in 2000, 33% scored Basic, and 38% scored below the Basic level.

NAEP Studies

***Comparisons Between KCCT and NAEP: Assessment Frameworks, Item Format, Item Content, Test Administration, Scoring, and Reporting*,** November 2003. HumRRO

Purpose

The No Child Left Behind Act of 2001 requires states benefiting from Title I grants to participate in NAEP. Results are to be considered by the U. S. Department of Education in its formal peer review of state accountability systems. The question of similarities and differences between the KCCT and NAEP arises in this context.

Audience

Department of Education leadership and Kentucky Board of Education: These parties and other groups will be interested in the relationship between the two assessments in the context of federal review considerations.

Methodology

Using released NAEP and KCCT items and rubrics, researchers elicited judgments from a convenience sample of 14 Kentucky teachers who had served on the Content Advisory Committee. Teachers' judgments concerned similarities and differences between NAEP content area frameworks and the Kentucky Core Content for Assessment, NAEP achievement levels and Kentucky Performance Standards, item format and content on grade 4 reading and grade 8 mathematics tests. The researchers themselves examined differences in test administration and scoring. Note that, due to the use of released rather than operational items on both tests, results should not be generalized to the operational KCCT and NAEP assessments.

Findings

- Examination of Kentucky Core Content for Assessment (KCCA) standards and NAEP frameworks suggest that the level of overlap between them is far from complete (see table below). Note, however, that respondents reported uncertainty with respect to 8% to 31% of the standards. Researchers indicate that in reading "…Kentucky's Core Content for Assessment represents a broader set of curricular topics than do NAEP standards (p. 4.)" However,"… the breadth of curricular topics is more similar in math than in reading (p. 6)."

Average Percentage of Observed Test Content Standards Matches by Direction		
Content/Grade	KCCT to NAEP	NAEP to KCCT
Reading		
4th	.26	.49
7th	.61	.88
Mathematics		
5th	.58	.89
8th	.61	.59
See HumRRO Report No. DFR-03-88, pp. 5 – 6		

- The researchers also report that that in a sort task, more NAEP reading items were "…placed at the higher levels of the taxonomy…(p. 52)," as a result of the fact that NAEP does not include a reading skills component.

- NAEP has short constructed response items and extended constructed response items while the KCCT has only extended constructed response items (i.e., open-response items).

- Kentucky open-response items are written to provide a structural guide to the students, whereas the NAEP items do not.

- NAEP is timed, whereas the KCCT is not timed.

Validity Evidence Related to Consequences of Test Use

Comparative Study of Standards and Indicators for School Improvement (SISI)and Academic Index for Selected Elementary Schools, HumRRO, 2004

Purpose (Why do the research?)

The purpose of the Commonwealth Accountability Testing System is to drive school improvement thereby, encouraging high student achievement. CATS consequences, such as, the reporting of results, recognition, information, and assistance, are intended to motivate, inform, and improve schools. The validation of KCCT results, therefore, must include evidence as to their consequences. The questions addressed in this study are: Does CATS fulfill its purpose, i.e., driving school improvement and increasing student achievement?

Audience

All constituencies interested in the educational success of Kentucky schools and students.

Background and Methodology

In biennial years all Kentucky schools are evaluated on the basis of their accountability indices and assigned scores: Meeting Goal, Progressing, and Assistance. Schools in Assistance are ranked and assigned to Levels 1, 2, and 3 (high to low). Consequences accrue to schools in Assistance. Those schools that score in the lowest third are audited by the KDE, Office of School Improvement, while those scoring in Levels 1 and 2 are reviewed. (The difference between a review and an audit lies in the composition of the team that does the work, criteria for review teams being somewhat more flexible than for audit teams.) In addition to auditing and reviewing schools in assistance, the Office of School Improvement reviews a number of schools classified as "Meets Goal." This provides a useful comparison group.

Both audits and reviews are conducted by teams of trained individuals who use the *School Improvement Standards and Indicators for School Improvement* (SISI), an evaluation tool featuring nine standards and 88 indicators organized under three headings: Academic Performance, Learning Environment, and Efficiency. The following indicators, 3.1.a and 3.1.b, are examples of Academic Performance Standard 3, Instruction:

- 3.1.a There is evidence that effective and varied instructional strategies are used in all classrooms;
- 3.1.b Instructional strategies and learning activities are aligned with the district, school, and state learning goals and assessment expectations for student learning.

Developed on the basis of extensive research and consultation on the part of the Office of School Improvement, the SISI define the elements of whole school improvement at the elementary, middle and high school levels. It is believed that these elements lead to effective schools. The authors of the present study correctly note, however, that inter-rater reliability with respect to ratings on the SISI has not been studied.

Audit and review teams visit schools to observe instruction and other activities and to talk frankly with faculty, administration, students, and parents. They collect information and rate each indicator on the SISI, using the four-category rating scale:

(1) Little or no implementation
(2) Limited development and partial implementation
(3) Fully functioning and operational level of development and implementation
(4) Exemplary level of development and implementation

Data collected on the basis of the SISI audits conducted from 1999 to 2003 were provided to the authors by KDE. Out of 188 elementary school cases, 144 featured complete data in both grades 4 and 5, and were, therefore, considered for the analysis. The authors grouped the school data on two dimensions: Biennial Accountability Index and Annual Academic Index. Recall that the Biennial Accountability Index is a two-year average including test data as well as non-academic data such as attendance, retention, and drop-out. The academic index is a weighted average of KCCT results across school

grade levels in a given year. The Accountability Index data were separated into five school categories: Assistance Levels 3, 2, and 1 (from low to high), Progressing, and Meets Goal. The Academic Index data were rank-ordered and separated by fifths. This procedure allowed the data to be arranged in a five-by-five table (see adaptation below). The authors selected cells (of schools) from this table, in such a way as to maximize potential contrasts in school ratings. Schools in the three corner cells of the table were selected for analysis:

(1) 47 schools ranked in the lowest fifth on the Academic Index *and* classified Accountability Level 3;
(2) 19 schools in the lowest fifth on the Academic Index classified as "Meets Goal;" and
(3) 12 schools ranked in the highest fifth on the Academic Index and classified as "Meets Goal."

This resulted in 77 elementary schools to be included in the analysis.

Number of Audited or Reviewed Kentucky Elementary Schools, 1999 - 2003 by Academic Index Group and Accountability Index Classification in Audit Year

Academic Index Group	Accountability Index Classification in Audit Year					
	Level 3	Level 2	Level 1	Progressing	Meets Goal	Number
Highest Fifth			1		11	12
			3	1	10	14
Middle Fifth			9	2	9	20
		5	6	10	9	30
Lowest Fifth	47	14		5	19	85
Total	47	19	19	18	58	161

Findings

(1) All Schools Need to Improve

- Level 3 Schools Scoring in Lowest Fifth: 50% rated SISI Category 1 or 2 on 99% of indicators
- Meeting Goal Schools Scoring in Lowest Fifth: 50% rated SISI Category 1 or 2 on 78% of indicators
- Meeting Goal Schools Scoring in Highest Fifth: 50% rated SISI Category 1 or 2 on 35% of indicators

(2) Significant between-group differences were detected using Kruskal-Wallis (K-W) and Mann-Whitney (M-W). Four statistical tests were conducted for each SISI indicator, each at 95% confidence level, resulting in family-wise alpha of approximately .20. Significant differences were demonstrated on 87 of the 88 SISI

indicators via the K-W. (The same 87 differences were demonstrated as statistically significant by Mann-Whitney comparison of the lowest and highest groups, i.e., Level 3 schools scoring in the lowest fifth as compared to the Meets Goal schools scoring in the highest fifth.)

(3) On fifteen indicators, the Mann Whitney U demonstrated significant differences between each of the two groups of schools that had met their goals as compared to the lowest group (lowest fifth level 3); these differences did not appear however, when the two groups that had met their goals were compared.

(4) On seven indicators, Mann-Whitney U demonstrated significant differences when the top group was compared to each of the two lowest-fifth groups, but no difference emerged between the two lowest-fifth groups. These 7 indicators were:

- 1.1.a – There is evidence that the curriculum is aligned with *Academic Expectations*, *Core Content for Assessment*, *Transformations*, and the *Program of Studies*.
- 1.1.e – The school curriculum provides specific links to continuing education, life, and career options.
- 2.1.e – Multiple assessments are specifically designed to provide meaningful feedback on student learning for instructional purposes.
- 4.1.c – Teachers hold high expectations for all students academically and behaviorally, and this is evidenced in their practice.
- 4.1.g – Teachers communicate regularly with families about individual students' progress.
- 5.1.c – The school/district provides organizational structures and supports instructional practices to reduce barriers to learning.
- 8.1.b – The master class schedule reflects all students have access to all of the curriculum.

While significant differences are observed between the two cell representing the lowest vs that representing the highest performing schools on 87 of 88 indicators, fewer significant differences (59 of 88) are demonstrated between the two lower groups of schools. The table presented in the Executive Summary, p. iii, summarizes results of the statistical tests. Readers will note the emergence of an interesting pattern with respect to these two lower groups (see shaded cells in the first column under "Mann-Whitney Test"). On Standards 6 (Professional Growth, Development, and Evaluation), significant differences are observed on 11of 12 Standards and Indicators; on Standard 7 (Leadership) significant differences are observed on 10 of 11 Standards and Indicators; on Standard 9 (Comprehensive and Effective Planning), 16 of 16 significant differences are observed. By contrast, significant differences on these indicators appear in lower numbers between the two groups of schools that meet their goals: Standard 6 (Professional Growth, Development, and Evaluation), 4 of 12; Standard 7 (Leadership), 3 of 11; Standard 9 (Comprehensive and Effective Planning), 2 of 16.

Recommendations

The authors suggest that further analyses include examination of qualitative information that is collected by audit/review teams in support the SISI ratings. This information could prove helpful in interpreting the quantitative results.

8. Other Studies

Additional topics addressed by KDE's research efforts include the relationship between students' test scores (KCCT and ACT) and their high school grades; trends shown on KCCT among disaggregated groups; the relationship between amount of testing and use of tests; and computer administration of the Kentucky Core Content Tests. Titles of these reports and abstracts provided by the authors appear below.

Relationships among Kentucky's Core Content Test, ACT Scores, and Students' Self-Reported High School Grades for the Classes of 2000 Through 2002, HumRRO, 2003

Abstract

As a part of Kentucky's ongoing examination of the validity and reliability of the Kentucky Core Content Test (KCCT), a major component of the Commonwealth Accountability Testing System (CATS), KCCT scores were compared with ACT scores for the period from 1999-2002. This report updates a similar study conducted by Hoffman (1998) comparing ACT scores with scores from KCCT's predecessor, the Kentucky Instructional Results Information System (KIRIS). Results were much the same as found during the earlier study. KCCT scores are correlated with ACT scores at the student and school level. In addition, students' self-reported grades and number of courses in mathematics and science are correlated with ACT and KCCT scores. Correlations between same-subject tests typically ranged from 0.50 to 0.65, indicating that while the different measures are related, they are not so highly related as to indicate that they are testing essentially the same set of content and skills. They are within the "Goldilocks" range, or not so high that they indicate that the tests do not have important differences, but not so low as to indicate that they measure entirely different content.

Kentucky students' KCCT scores have improved steadily over the three years studied while ACT scores, with the exception of mathematics, have declined slightly. Part of the analyses conducted as part of this study examined this pattern. In addition to student-level correlations, school-level correlations were also positive, indicating that if a school's mean score on KCCT was high, its mean ACT score was also high. When change in score was analyzed, a smaller correlation, which is expected from analysis of change scores, emerged. In all cases, this correlation was positive. However, the decline was small, and for schools that improved a great deal on KCCT it was smaller than for schools with smaller increases, allowing the positive correlation to emerge. So, although not immediately obvious, the data does not support the idea that the two tests represent divergent content. More simply, preparing students to do well on KCCT does not preclude them from doing as well on the ACT; in fact, the opposite is true. Schools that gained on KCCT had smaller losses, or even posted gains, on ACT.

Trends Between Student Gender, SES, Ethnicity, LEP, Disabilities and Kentucky Core Content Test Performance, HumRRO, 2003

Abstract

As part of recent "No Child Left Behind" (NCLB) legislation, education communities have been mandated to close gaps in academic performance across a range of student subgroups, including gender, ethnicity, disability, socioeconomic status (SES) and limited English proficiency (LEP). This report looks at mean scale score differences on the Kentucky Core Content Test (KCCT) among these five student subgroups. Mean scale scores are presented graphically to depict changes in performance by subgroup between testing years 1999-2002.

Student subgroup populations have remained stable over the four-year period, with no large fluctuations in size. Performance gaps between the various groups reflect expected patterns (based on other state and national measures of achievement), and these gaps have
been largely maintained over time. White students' mean scores are consistently higher than African-Americans' and Hispanics' mean scores, females higher than males (with the exception of science where males and females scores are essentially identical), students without disabilities higher than students with disabilities, students ineligible for free/reduced lunch higher than those meeting eligibility requirements (a proxy for SES), and non-LEP students higher than those with limited proficiency in English. An important caveat to these findings, however, is that the variability within any of the analyzed groups is much larger than the difference in their mean scores. Graphics in this report include bars at each data point depicting one standard deviation above and below the mean, which will encompass roughly the two-thirds of students making up the center of the overall group distribution. Membership in a traditionally lower-achieving group does not indicate that particular students in the group will be lower achieving; conversely,
many students in lower scoring groups score above the mean of the higher scoring group. While plotting means highlights differences between groups, adding the +/-1 standard deviation ranges highlights the groups' substantial overlap.

In addition to the comparison of means, multiple regression analysis was conducted in order to further explore the relationship between subgroup membership and KCCT performance. In nearly all instances, students' subgroup membership added to the prediction of their KCCT scale score, and the direction of these relationships reflected patterns depicted graphically. Students' gender, ethnic, disability and socioeconomic status do have an impact on performance on assessments such as the KCCT. This paper calls for further exploration of these gaps in student performance through the replication and expansion of these analyses for subsequent testing years.

Survey of the Amount of Student Testing and the Uses of those Tests, HumRRO, 2004.

Abstract

This report investigates critics' claims that Kentucky's students are subjected to an overabundance of testing, which thereby detracts from instruction and consequently student learning. To investigate the validity of this claim, Kentucky's district assessment coordinators (DACs) were surveyed about their district's assessment programs. Survey results indicate that the majority of districts administer three or fewer tests in addition to those required by the Commonwealth Accountability Testing System (CATS). The majority of these non-CATS mandated tests are administered at the early elementary school level; grades not currently assessed by any of the CATS component tests. DACS indicated that more time was spent on test preparation activities than test follow-up activities, particularly for Kentucky Core Content Tests (KCCT). In addition, DACs generally indicate that: (1) their current assessment program elicits moderate to high levels of pressure among students and teachers, (2) the current amount of testing represents an assessment system that equals or includes more than their ideal program, and (3) the costs of the current assessment program equal or outweigh the benefits. Also, when asked which tests should be retained or added, DACs frequently indicated that they would like to see CTBS tests for all school levels. More than half the respondents also favored keeping the KCCT exams at all grade levels, but the responses were not as positive as for CTBS tests. The implications of these results are discussed in relation to the augmented off-grade testing scheduled to begin in spring 2005.

CATS Online: Logistic and Construct Evaluation of Computer Administered Assessment, HumRRO, 2003

Abstract

In an effort to meet the needs of Kentucky students who have traditionally required a human reader as part of their testing accommodations, the Kentucky Department of Education (KDE), in partnership with eCollege[SM], developed the CATS Online Assessment. The CATS Online Assessment allows students to read the Kentucky Core Content Test (KCCT) from a computer screen, using either text reader or screen reader software. With text/screen reader software, students highlight a portion of text that they wish to read and the computer reads aloud to them.

This report examines two major issues regarding the use of computer-based assessment as a valid measure of student performance. First, the report discusses whether the CATS Online Assessment allows students to demonstrate their content knowledge in a way similar to their daily classroom activities. Second, the report explores logistical issues associated with the administration of the test.

Although text/screen reader technology software is a fairly new phenomenon at most Kentucky schools, most participating teachers and students had at least one year of experience working with the technology. Though initially concerned about technical issues, students and teachers seemed generally comfortable navigating the assessment, and students largely reported preferring the computer-based test to the way they have taken it in the past. Similarly, most test proctors felt that the online test helped rather than hindered students in their understanding of test content. Several logistical issues emerged, ranging from technical to procedural concerns. However, all participating students were able to complete the assessment, and no student data were lost. The CATS Online Assessment appears to be allowing students to demonstrate content knowledge in a way to which they are accustomed, with minimal logistical problems.